Impact of microphone array configurations on robust indirect 3D source localization

IMPACT OF MICROPHONE ARRAY CONFIGURATIONS ON ROBUST INDIRECT 3D ACOUSTIC SOURCE LOCALIZATION

Elizabeth Vargas, Keith Brown

Heriot-Watt University
Edinburgh, United Kingdom

Kartic Subr

University of Edinburgh
Edinburgh, United Kingdom

ABSTRACT

Acoustic source localization (ASL) is an important problem. Despite much attention over the past few decades, rapid and robust ASL still remains elusive. A popular approach is to use a circular array of microphones to record the acoustic signal followed by some form of optimization to deduce the most likely location of the source. In this paper, we study the impact of the configuration of microphones on the accuracy of localization. We perform experiments using simulation as well as real measurements using a $72-$microphone acoustic camera which confirm that circular configurations lead to higher localization error, than spiral and wheel configurations when considering large regions of space. Moreover, the configuration of choice is intricately tied to the optimization scheme. We show that direct optimization of well known formulations for ASL yield errors similar to the state of the art (steered response power) with $6\times$ less computation.

Index Terms— 3D acoustic source localization, microphone array configuration

1. INTRODUCTION

The problem of estimating the 3D position of objects is called *localization*. There has been tremendous advancement in robust localisation of objects using visual features. The use of audio sensing has important advantages such as reliability under poor illumination conditions, relatively inexpensive sensing equipment and the prevalence of signal processing (1D) tools. There have been attempts to use audio localization examples include: in the automotive industry [1], in robotics [2] and in scene understanding [3]. *Acoustic source localization* (ASL) is typically achieved by leveraging known discrepancies in measurements of the emitted signal at multiple locations. ASL algorithms may exploit differences in time, amplitude or in both time and amplitude.

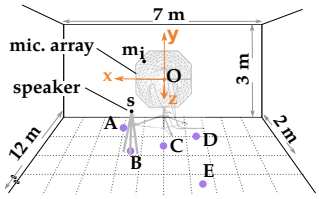
Some approaches to ASL, such as steered response power [4, 5], directly solve for the most likely position of the source amongst a grid of candidate locations. ''Indirect" methods, on the other hand, first estimate the times of arrival (TOA) at the sensors (microphones) or time differences of arrival (TDOA) across pairs of microphones and then use this information to deduce the most likely position of the source via multilateration [6, 7]. Although indirect methods are simpler to express as a least squares optimization [8], the resulting objective function is non-convex and often does not lend itself to analytical solution. Various reformulations of these using weighted least squares, linear correction least squares, constrained least squares. convex constrained least squares [9], total weighted least squares [10] and weight constrained total least squares [11] have been analyzed. Direct methods are believed to be more robust to noise and reverberation [4].

A uniform circular array of microphones[12, 13] along with a ring configuration [14] is a common choice for taking measurements since azimuthal angles to sources are considered more important than elevation. The advantage of *acoustic cameras* with such arrays is that they can focus on specific targets [15, 16], which is useful for speech processing. The resolution in elevation has recently shown to be improved by using a 2.5D circular array [17]. While there have been a few results on the use of spherical arrays, multiple spheres [18], randomly placed microphones [19, 20] and spiral configurations [21], there is surprisingly little analysis of the impact of the geometric structure of the measurement array on particular optimization algorithms for ASL.

In this paper, we adopt an optimization (sequential least squares programming) approach for indirect ASL. We focus on the core problem of localizing a single source. Other work towards estimating TDOA for multiple sources are directly applicable. Although the objective function we choose is nonlinear and non-convex, we show using *simulation as well as real data* that the method is robust to noise and reverberation. Our experiments verify that it is comparable to SRP for real data while being $6\times$ more efficient to compute. Using this optimization scheme, we study the localization error resulting from different geometric structure for the microphone array. Our results show that circular arrays produce the highest errors (across space) and are therefore least desirable.

2. OBJECTIVE FUNCTION AND OPTIMIZATION

Consider a source at location $\mathbf{s}$ that emits an acoustic signal at some arbitrary time t^*. Let the measurements of the emitted sound be recorded by an array of M microphones located at



(a) Setup

signal	SRP		TDOA (100)		TDOA (all)	
	Rel. Err %	Time in min	Rel. Err %	Time in min	Rel. Err %	Time in min
chirp	14.7 (25.9)	3 (0.2)	14.2 (25.9)	0.5 (0.01)	12.1 (23.2)	4.5 (0.03)
gunshot	11.0 (13.3)	2.58 (0.2)	9.6 (12.8)	0.4 (0.02)	6.4 (3.5)	2.4 (0.02)
dogbark	16.0 (28.5)	2.49 (0.1)	58.9 (38.8)	0.4 (0.02)	48.5 (44.6)	2.4 (0.02)
speech	13.2 (21.1)	2.63 (0.1)	15.2 (23.5)	0.4 (0.02)	12.9 (22.5)	2.5 (0.02)

(b) Errors and computation time comparison across all microphone configurations

Fig. 1. (a) Our setup and coordinate system. (b) Table comparing errors and time for SRP with TDOA optimization using 100 of the C_2^{72} mic pairs (middle) and using all pairs. Standard deviations are shown within parantheses.

$\mathbf{m}_i,\ i = 1, 2, ..., M$ and the times taken by the signal to travel from $\mathbf{s}$ to $\mathbf{m}_i$ be t_i. If the distance between the source and the i^{th} microphone is $d_i \equiv \|\mathbf{m}_i - \mathbf{s}\|$, then $t_i = d_i/c + t^*$ where c is the speed of sound in air and t^* is not generally known.

Time of arrival In the case that the times of arrival at the microphones are measured as $\tilde{t}_i$, we pose the ASL problem as one of jointly determining $\mathbf{s}$ and t^* as

$$O_1 : \quad \arg\min_{\mathbf{s},t^*} \sqrt{\sum_{i=1}^{M} (\tilde{t}_i - t_i)^2} \tag{1}$$

Time Difference of Arrival (TDOA) Another possibility is to note the difference in measured times between a pair of microphones, $\tilde{\tau}_{ij} \equiv \tilde{t}_i - \tilde{t}_j$, or TDOA. The literature is rich with methods to estimate TDOA. We choose the popular Generalized Cross-Correlation Phase Transform (GCC-PHAT) [22]. Then, we perform ASL by optimizing [8]:

$$O_2 : \quad \arg\min_{\mathbf{s}} \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{M} \left(\tilde{\tau}_{ij} - \tau_{ij}\right)^2}, \tag{2}$$

where $\tau_{ij} = (t_i - t_j)$.

For both formulations O_1 and O_2, since we know that the solution is constrained by the dimensions of the room, we supply these constraints as linear inequalities. We solve the constrained non-linear optimization using Sequential Least Squares Programming (SLSQP) which is an interative procedure. In each iteration, a constrained quadratic programming sub-problem is constructed so that the chain of solutions converges to a local mininum [23]. Each subproblem replaces the objective function with a local, quadratic approximation subject to local affine approximations of the constraints. We used a Broyden-Fletcher-Goldfarb-Shanno (BFGS) approximation to update the Hessian matrix required for the local quadratic approximation and chose the step length using an L_1 test function. The optimizer used to solve each subproblem is a modified version of NNLS [24]. We used the following parameters as inputs to the optimizer: iterations = 1500, accuracy = 1e-20, epsilon = 1.49e-08.

2.1. Experiments

We performed experiments using simulation as well as real measurements using an *gfai tech AC_Pro Acoustic Camera system* consisting of 72 reconfigurable microphones sampling at 192kHz. We used three different microphone configurations: ring, wheel and spiral. Using each configuration, we measured recorded sounds played by a *Bose Soundlink Bluetooth Mobile Speaker II, Model 404600* in five different calibrated positions within a room of size $12m \times 7m \times 3m$. The speaker was positioned, using a tripod, to be on the plane $y = -0.32$ for all five positions A, B, C, D and E. For each position we acquired three recordings. Fig. 1 illustrates the setup. We repeated the experiments for 4 different audio signals [25]: chirp, gunshot, dogbark and speech.

Simulation: noisy TOA and TDOA We tested the robustness of the proposed optimization by evaluating the relative error in localization for different simulated degrees of noise σ in the estimated TOA and TDOA values. To enable comparison across multiple sources locations, we express σ for each source location as a percentage of the time taken for sound to travel from $\mathbf{s}$ to the center of the microphone array $\mathbf{O}$. We use a Gaussian model for the noise in simulated TOA $\tilde{t}_i = t_i + \eta$ and for TDOA $\tilde{\tau}_{ij} = \tau + \eta$ where

$$\eta \sim \mathcal{N}\left(0, \frac{\sigma}{100}\frac{\|\mathbf{s} - \mathbf{O}\|}{c}\right). \tag{3}$$

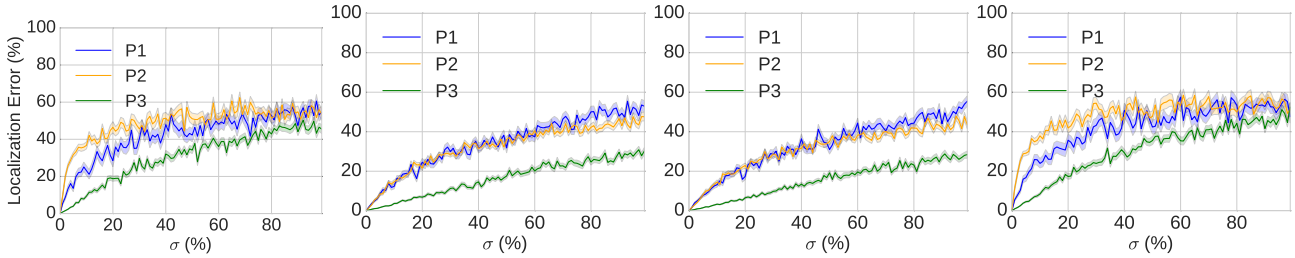
We measure relative error, expressed as a percentage of the distance from the source to the camera, as the evaluation metric for the accuracy of localization:

$$\text{error}(\%) = \frac{\|\mathbf{s} - \tilde{\mathbf{s}}\|}{\|\mathbf{s} - \mathbf{O}\|} * 100, \tag{4}$$

where $\tilde{\mathbf{s}}$ is the source location estimated by the optimization.

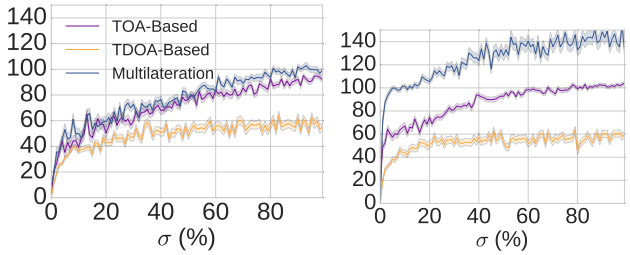
We compared optimizations for TOA and TDOA with multilateration [7]. Fig. 3 depicts plots of relative localization error (Y-axis) as the noise in the simulation is increased (X-axis). We performed two versions of the experiment: First, assuming that microphone and the sound source are synchronised ($t^* = 0$ in Fig. 3a), then without that assumption by setting $t^* = 0.01s$.

Simulation: microphone configuration We estimated the localization error at different points in space. Since it would be rather tedious to repeat real measurements over a dense grid of source locations, we obtained this via simulation. For each source position on a dense grid, we estimated the localization errors for three microphone configurations. The three configurations were identical to those used for real measurements with our acoustic camera, consisting of 72



(b) Ring (c) Wheel (b) Spiral (c) Random

Fig. 2. The plots show simulated relative localization error (Y-axes) for increasing degrees of noise (X-axes) observed at three source locations: P1: (-2,-1,4), P2: (-1,0.5,3), P3: (0.4,0.7,1.05).



(a) synchronized ($t^* = 0$) (b) unknown t^* (10ms)

Fig. 3. Plots comparing relative localization errors using O_1 (TOA), O_2 (TDOA) and multilateration [7] (a) when the speaker is synchronized with microphones and (b) when the time of emission is unknown.

microphones. Each configuration results in different TOA and TDOA values, due to the different microphone positions. When noise is added to these TOA and TDOA values, each configuration reveals a characteristic heat-map for localization error over space. Fig. 4 visualizes these heatmaps for $\sigma = 100\%$ simulated error, along with the corresponding error histograms. The errors were averaged over 100 trials for each grid point. We chose a grid over $x = [-2, 2]$, $z = [0, 4]$ and $y = -0.32$, with a resolution of 10 cm, so that it matches our experiments with real data. For three positions $P1 \equiv (-2, -1, 4)$, $P2 \equiv (-1, 0.5, 3)$ and $P3 \equiv (0.4, 0.7, 1.05)$, we plotted error as a function of noise for four different microphone configurations (fig. 2).

Real Data: Comparison with SRP We used optimization scheme O_2 to localize a speaker placed in five positions $A \equiv (2.0, -0.32, 0.5)$, $B \equiv (1.5, -0.32, 2.0)$, $C \equiv (0.0, -0.32, 1.5)$, $D \equiv (-1.5, -0.32, 1.0)$ and $E \equiv (-1.5, -0.32, 3.5)$. Fig. 5 plots relative errors (Y-axes) for three different microphone configurations (X-axes) at the chosen five locations (columns). The three rows of plots correspond to results obtained using SLSQP, SRP and Bayesian optimization respectively. Errorbars (standard deviation) are shown with black lines on top of the bars.

2.2. Results and discussion

Microphone configurations Our results overwhelmingly suggests that circular (ring configuration) arrays are worse than spiral or wheel configurations when considering relative localization error over a wide range of positions. Although slightly exaggerated (100% noise), our simulation results (fig. 4) show regions (top view) that are error prone when using circular arrays. This is also true for our real measurements (fig. 5), where the results obtained for position C are worse for ring than for wheel or spiral using any of the three localization techniques. The yellow bars in the first row show that the errors observed with real data correspond to errors obtained with about 10% noise in our simulation.

Comparison with multilateration Our experiments revealed that both optimization strategies O_1 and O_2 result in lower relative errors than state of the art multilateration [7]. This is particularly true when the time of emission of the signal is unknown and when the emitter is not synchronized with the microphones ($t^* \neq 0$). When $t^* = 0$, we observed that our implementation of the multilateration algorithm has similar accuracy to optimizing O_1 (TOA). Our proposed approach to optimizing O_2 (TDOA) has the least relative errors and remains unaffected by t^*.

Comparison with SRP A common criticism that is faced by indirect methods is that the optimization is not as robust as direct methods such as SRP. However, our results (Table 1) show that our localization error is comparable to SRP but is more efficient. For this comparison, we used an efficient implementation of SRP that leverages stochastic region contraction [5] and a naïve implementation of our optimization in python. Just as with their method, the accuracy of the proposed optimization may also be traded for performance.

Accuracy vs performance One way to approximate the localization is to modify the nested summation in O_2 to only consider some of the microphone pairs. We studied convergence plots of localization error for different source positions, as the number of microphone pairs is increased from just 1 pair to all pairs (C_2^{72}). We observed that the error generally drops below 10% for 100 mic pairs (see Table 1 for the corresponding computation times), except for the dogbark signal. Figure 6a plots relative error averaged across spatial locations for all four test signals using only 100 microphone pairs.

Bayesian optimization We tested a Bayesian optimizer with O_2 as its loss function ($\kappa = 1$). This took an order of magni-

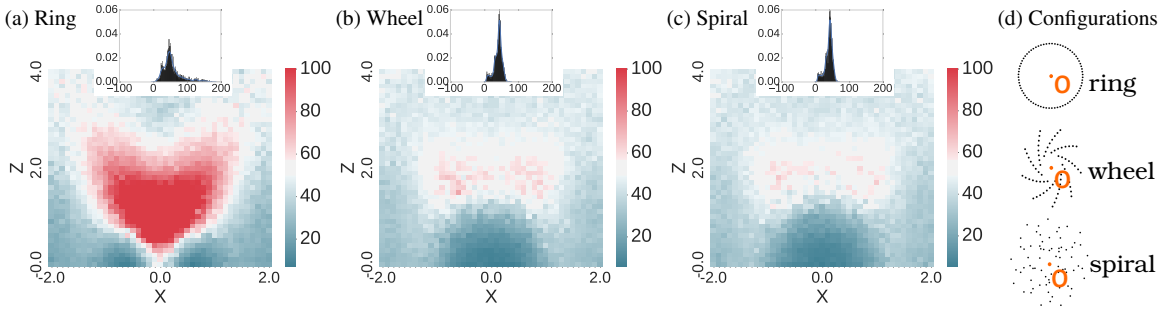
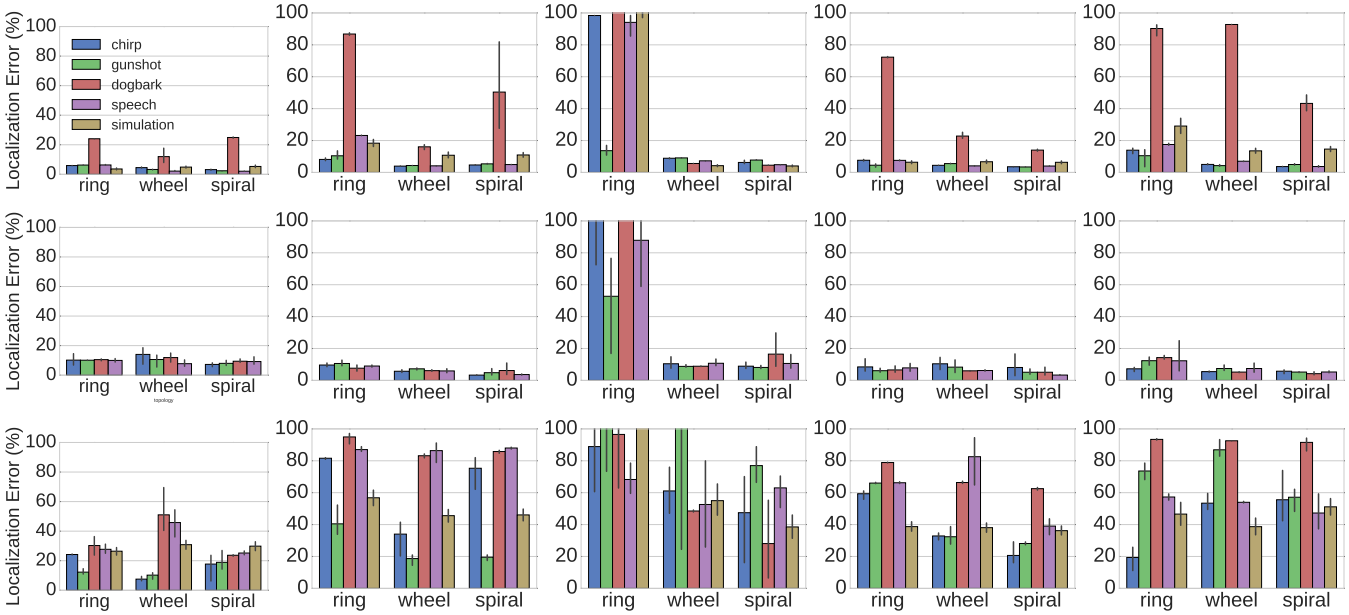


Fig. 4. Relative error percentages visualized as heatmaps obtained using simulations, at 100% noise, for a $2\mathrm{m} \times 2\mathrm{m}$ room. 100 estimates were averaged for the error estimate at each grid position. The insets show the distributions of errors as histograms.

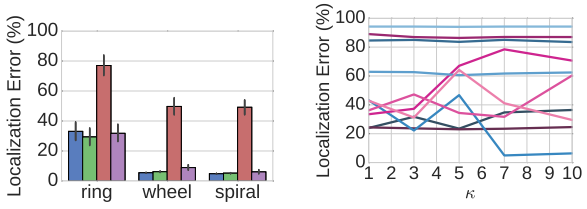


(a) A: (2.0,-0.32,0.5) (b) B: (1.5,-0.32,2.0) (c) C: (0.0,-0.32,1.5) (d) D: (-1.5,-0.32,1.0) (e) E: (-1.5,-0.32,3.5)

Fig. 5. Localization Error using SQLP and simulation with 10% noise (top row) SRP (2nd row) and Bayesian Optimization with exploitation (3rd row)

tude longer than SQLSP and the resulting errors were larger. We tested with various degrees of the κ parameter to trade-off exploitation versus exploration, speculating that the poor performance was due to the presence of multiple local minima. However, the plot (fig. 6b) shows that exploitation ($\kappa = 1$) performs better than exploration ($\kappa = 10$) in most cases. The number of iterations and tolerance were set so that optimizer converged to the reported solutions, suggesting that the problem is not due to multiple local minima.

Limitation One of the drawbacks of indirect localization achieved by minimizing O_2 is its dependency on the estimated TDOA values. Although our results show that GCC-PHAT is accurate enough to yield localization errors comparable to SRP, the former performs worse when dealing with signals with repeating patterns such as the barking of a dog (red bar in fig. 5). Interestingly, our localization was more robust to reverberation (when the source was placed at room boundaries) than to repetitive macro-structures. Perhaps using full signal correlation matrices, as adopted by spectral estimation techniques, would resolve this problem.



(a) Using few mic pairs (b) Bayesian Optimization

Fig. 6. (a) Errors (real data) for four signals (colored bars) across spatial locations using 100 mic. pairs. (b) Exploitation ($\kappa = 1$) vs exploration ($\kappa = 10$) for dogbark (blue) and speech (purple) for spiral configuration.

3. REFERENCES

[1] José A Ballesteros, Ennes Sarradj, Marcos D Fernández, Thomas Geyer, and M[a] Jesús Ballesteros, "Noise source identification with beamforming in the pass-by of a car," *Applied Acoustics*, vol. 93, pp. 106–119, 2015.

[2] Ivan Marković and Ivan Petrović, "Speaker localization and tracking with a microphone array on a mobile robot using von mises distribution and particle filtering," *Robotics and Autonomous Systems*, vol. 58, no. 11, pp. 1185–1196, 2010.

[3] Ryosuke Kojima, Osamu Sugiyama, and Kazuhiro Nakadai, "Scene understanding based on sound and text information for a cooking support robot," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2015, pp. 665–674.

[4] Markus VS Lima, Wallace A Martins, Leonardo O Nunes, Luiz WP Biscainho, Tadeu N Ferreira, Maurício VM Costa, and Bowon Lee, "A volumetric srp with refinement step for sound source localization," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1098–1102, 2015.

[5] Hoang Do, Harvey F Silverman, and Ying Yu, "A real-time srp-phat source location implementation using stochastic region contraction (src) on a large-aperture microphone array," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 1, pp. I–121.

[6] Yiteng Huang, Jacob Benesty, Gary W Elko, and Russell M Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[7] Orhan Oçal, Ivan Dokmanic, and Martin Vetterli, "Source localization and tracking in non-convex rooms," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. Ieee, 2014, pp. 1429–1433.

[8] Jacob Benesty, M Mohan Sondhi, and Yiteng Huang, *Springer handbook of speech processing*, Springer Science & Business Media, 2007.

[9] Xiaomei Qu and Lihua Xie, "An efficient convex constrained weighted least squares source localization algorithm based on tdoa measurements," *Signal Process.*, vol. 119, no. C, pp. 142–152, Feb. 2016.

[10] K. Yang, J. An, X. Bu, and G. Sun, "Constrained total least-squares location algorithm using time-difference-of-arrival measurements," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 3, pp. 1558–1562, March 2010.

[11] Cao Jing-min, Wei He-wen, and Yu Jian, *Weighted Constrained Total Least-Square Algorithm for Source Localization Using TDOA Measurements*, p. 739746, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[12] Despoina Pavlidi, Matthieu Puigt, Anthony Griffin, and Athanasios Mouchtaris, "Real-time multiple sound source localization using a circular microphone array based on single-source confidence measures," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 2625–2628.

[13] Despoina Pavlidi, Anthony Griffin, Matthieu Puigt, and Athanasios Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[14] Guillaume Lathoud, Jean-Marc Odobez, and Daniel Gatica-Perez, "Av16. 3: An audio-visual corpus for speaker localization and tracking.," in *MLMI*. Springer, 2004, pp. 182–195.

[15] Zebb Prime and Con Doolan, "A comparison of popular beamforming arrays," *Australian Acoustical Society AAS2013 Victor Harbor*, vol. 1, pp. 5, 2013.

[16] David Ayllón, Roberto Gil-Pita, Manuel Utrilla-Manso, and Manuel Rosa-Zurera, "An evolutionary algorithm to optimize the microphone array configuration for speech acquisition in vehicles," *Engineering Applications of Artificial Intelligence*, vol. 34, pp. 37–44, 2014.

[17] Mingsian R Bai, Chang-Sheng Lai, and Po-Chen Wu, "Localization and separation of acoustic sources by using a 2.5-dimensional circular microphone array," *The Journal of the Acoustical Society of America*, vol. 142, no. 1, pp. 286–297, 2017.

[18] X. Pan, H. Wang, F. Wang, and C. Song, "Multiple spherical arrays design for acoustic source localization," in *2016 Sensor Signal Processing for Defence (SSPD)*, Sept 2016, pp. 1–5.

[19] Mohammad J Taghizadeh, Saeid Haghighatshoar, Afsaneh Asaei, Philip N Garner, and Hervé Bourlard, "Robust microphone placement for source localization from noisy distance measurements," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2579–2583.

[20] Roberto Macho-Pedroso, Francisco Domingo-Perez, Jose Velasco, Cristina Losada-Gutierrez, and Javier Macias-Guarasa, "Optimal microphone placement for indoor acoustic localization using evolutionary optimization," in *Indoor Positioning and Indoor Navigation (IPIN), 2016 International Conference on*. IEEE, 2016, pp. 1–8.

[21] Chiong Lai, Sven Nordholm, and Yee-Hong Leung, "Design of robust steerable broadband beamformers with spiral arrays and the farrow filter structure," in *Proceedings of IWAENC 2010*. Ortra, 2010, vol. 90, pp. 653–669.

[22] Pasi Pertila, Matti S Hamalainen, and Mikael Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 11, pp. 2393–2402, 2013.

[23] Philip E Gill and Elizabeth Wong, "Sequential quadratic programming methods," in *Mixed integer nonlinear programming*, pp. 147–224. Springer, 2012.

[24] Dieter Kraft, "A software package for sequential quadratic programming," *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

[25] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.